International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Clinically Weighted Zero Trust: Securing Healthcare AI Where Lives Depend on It

Mostafa Rahmany

Email: *globalmostafa[at]gmail.com*

Abstract: ***Health systems continue to be a significant focus of ransomware; when large-scale attacks (e.g., the Change Healthcare ransomware breach in 2024 (Kannarkat, 2024; Alder, 2025; TechCrunch, 2025)) seemed to have a domino effect of disrupting claims, eligibility, and pharmacy services across the country, with an unprecedented impact on patients. Current defences are over-reliant on flat networks, VPN trust, and endpoint agents that are incompatible with regulated medical devices. The NHS WannaCry post-incident review (National Audit Office, 2017) and the HSE Ireland post-incident review (HSE,2022) confirm the lack of patching, segmentation, and response orchestration. Design and test ClinDefend-ZTA, a clinical-safety-first Zero-Trust and deception hospital architecture that: (i) minimises both lateral movement and care harm; (ii) isolates any vulnerable devices flagged by SBOM scanners; and (iii) can contain breaches faster than legacy tools, observably reducing time-to-contain (TTC) and service outage. The policy model is selected that assigns device criticality to authentication, authorization, and micro-segmentation strengths; introduces decoys (credentials, DICOM shares, HL7 test endpoints) at convergence points, and incorporates runtime SBOM gates to steer device traffic. To perform the said evaluation, it is based on a lab testbed, a retrospective what-if replay of three real breaches (NHS, 2017; HSE, 2021; U.S. 2024 clearinghouse event). Prototypes are expected to achieve TTC reduction by 50+ per cent, impacted subnet reduction by 40+ per cent, and zero clinical safety incidents during simulated lock-downs, and this qualifies as the performance norm compared to perimeter+EDR baselines. Hospitals can implement ClinDefend-ZTA with existing identity, MFA, NAC, and micro-segmentation products, in line with FDA device cybersecurity recommendations and HHS/HICP best practices, without having to rely on ML/AI products and therefore being impeded by their validation.***

Keywords: zero trust, hospital AI security, ransomware resilience, software bill of materials, patient safety

1. Introduction

1.1 Research Problem and Motivating Context

The swift digitalization of the healthcare industry has resulted in hospitals becoming data-intensive settings to be managed with the help of artificial intelligence (AI) regarding diagnosis and triage, work-related processes, and administrative efficiency. Although such a reorientation benefits patient care, it also leads to emerging complex cyber risks. The main difference is that the surge in common-level ransomware, as in cyberattacks on Universal Health Services in 2020 and Common Spirit Health in 2022(Powell, 2023; TechCrunch, 2022; Alder, 2023; Cybersecurity Dive, 2022), has also shown that cyberattacks not only disrupt hospital operations but can pose a direct threat to patient care by delaying emergency and critical care support. In hospitals, traditional security measures such as firewalls, endpoint detection, and partitioned networks are meant to secure infrastructure but do not address the security and privacy of AI models, training datasets, and inference pathways. A growing trend in incorporating AI in clinical decision-making establishes a critical need to implement cybersecurity frameworks that protect AI elements as the first-order entities. Without these frameworks, attackers can sabotage models, steal training data, or otherwise engineer inference results, disastrously impacting the patients.

1.2 Research Objectives, Scope, and Significance

The main goal of this study is to develop and assess a clinically weighted zero-trust Model supply chain and inference process with verifiable provenance, confidential inference, and patient-safety-oriented risk prioritization. It includes hospital-based applications of AI, such as radiology, triage systems, EHR analysis, and administrative automation. By focusing on securing vulnerabilities at the model level, in addition to network or device levels, this research aligns more closely with clinical realities that AI systems compromised at a model level directly impact patient outcomes. It is essential to create a framework that is resilient to cyber threats and prioritizes clinical harm risks: services that can cause the most significant clinical harm to others, such as emergency triage, must be secured before less harm-risky automation, such as claims processing. This two-fold emphasis on technical assurance and clinical prioritization sets the work apart from current cybersecurity paradigms.

1.3 Gap in Literature and Practice

A few publications on healthcare cybersecurity have been dedicated to ransomware resilience, network segmentation, and hardening of EHR systems. Related work on zero-trust networks in healthcare have suggested identity-based frameworks, though these have been left in theory, and do not consider specific threats to AI pipeline integrity (Corpuz, 2023). Likewise, the growing field of adversarial machine learning in medical imaging points to vulnerability to model poisoning and inference manipulation, However, most research focuses solely on algorithmic solutions, overlooking broader architectural mitigation. (Crigger et al., 2022). The intersection of zero-trust architecture, AI model provenance, confidential execution, and patient safety prioritisation is missing. No previous research unites these areas in an applicable, hospital-ready design. This study will close that gap, integrating model verification via attestation and Trusted Execution Environments (TEEs) into a clinically weighted zero-trust framework of healthcare AI.

1.4 Contributions and Novelty

This paper has four fundamental contributions.

1) It presents the idea of the Clinically Weighted Zero-Trust that would incorporate risk scoring around patient safety in hospital cybersecurity policy decisions.
2) It proposes a new model that federates in-toto attestations and SBOMs to deliver demonstrable provenance of AI models and datasets, which guarantee supply-chain provenance.
3) Confidential inference. It integrates confidential inference through TEEs to avoid unauthorised access to models and patient data during inference execution.
4) It provides a proof-of-concept implementation and comparative analysis that indicate superior resilience, reduced mean-time-to-detect/respond, and a favourable cost-benefit ratio compared to the current hospital cybersecurity solutions.

Collectively, these contributions lead healthcare to cybersecurity evolution past infrastructure-related defences toward patient-centric, AI-component-centric security. Innovation is the correlation of zero-trust principles with clinical harm weighting, which has not been considered in scholarly and medical practice before.

This study is significant because it bridges cybersecurity architecture with clinical risk prioritisation, offering a scalable framework for hospitals to mitigate AI-targeted cyber threats while protecting life-critical services. It also lays groundwork for regulatory and technological advancement in securing future health systems.

Table 1: Summary of Contributions and Novelty of CWiZ-AI

Contribution Area	State of the Art (2019–2024)	Novelty in CWiZ-AI Framework
Zero-Trust in Healthcare	Conceptual, focused on users/devices	Extends zero-trust to **AI components** (models, datasets, pipelines)
Adversarial ML Defences	Algorithmic countermeasures	**Provenance verification + runtime policy enforcement** beyond algorithm-level fixes
Supply Chain Transparency	FDA recommends SBOMs; partial adoption	Integration of **in-toto attestations + SBOMs** for verifiable AI provenance
Confidential Computing	Growing interest in TEEs; limited in healthcare AI	**Confidential inference for clinical AI** to protect PHI and model IP
Patient-Safety Prioritisation	Absent in existing cybersecurity frameworks	Introduction of **a patient–safety–weighted risk index** guiding containment
Real-World Validation	Case studies analysed post-incident (UHS, HSE, Common Spirit)	**Counterfactual analysis + PoC simulation**, showing quantifiable improvements

1.5 Roadmap of the Paper's Structure

The structure of the remainder of the paper is outlined below:

Section 3 describes the methodology and background. Section 4 introduces the proof-of-concept. Section 5 outlines real-life case studies. Sections 6 to 9 provide analysis, cost-benefit review, discussion, and conclusions, respectively.

2. Literature Review

2.1 Foundations of Healthcare Cybersecurity

Cyber-attacks on healthcare organisations have always been among the highest-value cyber targets because of the nature of the medical organisation services and the sensitivity of the data, which may contain patient information (Sendelj & Ognjanovic, 2022). Hospitals have been particularly at risk of ransomware attacks, which have been forcing them to suspend patient care and accumulate multimillion-dollar costs (Pham et al., 2024). A recent systematic review points out that healthcare is one of the least resilient critical infrastructure sectors in the face of cyberattacks, with aged IT systems, fractious governance, and inadequate workforce capacity as ongoing issues (Herrera et al., 2023). Intrusion detection systems, anti-malware, and firewalls offered a certain amount of protection but have repeatedly been unable to stop high-level attacks. This reflects a fundamental shortcoming, i.e., the current tactics are intended to protect stationary networks but fail to secure dynamic AI pipelines that process clinical data for critical decision-making

2.2 Zero-Trust Architecture in Healthcare

Zero-Trust Architecture (ZTA) has become a prominent cyberspace framework holding the slogan no-trust, always-verify concerning all users, devices, and workloads (Lund et al., 2024). In healthcare, it has been reported that ZTA has the possibility of overcoming insider threats and lateral movement as it can make trust boundaries more granular. however, the literature typically is conceptual with few accounts of ZTA implementation outside of pilot projects. In addition, most health-related research focuses on network and EHR security instead of AI pipelines (Neprash et al., 2022). This is noteworthy because, in the case of a patient, AI is promoting rapid radiology triage and predictable bed management. Hospitals will run the risk of applying an outdated security lens to a changing threat environment without adopting ZTA with model supply chain specificities and inference confidentiality.

2.3 Adversarial Machine Learning in Clinical Settings

Parallel to ZTA, adversarial machine learning (AML) has exposed deep flaws in medical AI's resilience to withstand attacks. Finding that subtle image perturbations in radiology images might confound model predictions and result in unsafe clinical decisions, Finlayson et al. (2019) showed that adding random noise to clean radiology images induces image perturbations that might affect prediction, termed image perturbations. The later literature has developed upon these risks, demonstrating that the process of training and inferencing could be poisoned, with the capability of an adversary to inject triggering content that is not detected

(Srivastava et al., 2025). More recently, Chinta et al. (2025) devoted an article to the fact that medical AI systems are catalogues of exceptions since their extreme sensitivity to information makes them appealing targets of information theft and manipulation. Despite research in technical solutions like adversarial training and differential privacy, such approaches are inadequate in a hospital context, where provenance verification, runtime monitoring, and prioritisation of clinical risk are also required.

2.4 Software Supply Chain Security and Attestations

The 2020 SolarWinds intrusion raised concerns about supply chain security in software, which sparked efforts to improve provenance verification using Software Bill of Materials (SBOMs) and attestation schemes. In the health sector, the U.S. Food and Drug Administration (FDA) has issued healthcare-specific guidance to use SBOMs in medical device cybersecurity (Health, 2022). Some scholarly reviews state that SBOM adoption is a good practice to enhance transparency, but it is not enough unless it is continuously verified during deployment (Carmody et al., 2021). End-to-end build, test, and deployment attestation. There have been proposed tools, such as in-toto, that allow attestation of the whole build, test, and deployment stages (Leligou et al., 2024). There is also a lack of such methods in healthcare AI pipelines in the literature. Lacking model-level provenance creates a vulnerability by allowing malicious models into registries that malicious actors may tamper with and add untrusted models into registries or even poison them. This again adds to the necessity of developing an attestation-based zero-trust solution that could be optimised for clinical AI.

2.5 Confidential Computing and Trusted Execution Environments

Trusted Execution Environments (TEEs) such as secure computing are beginning to play the critical role of protecting data in use by encrypting computations in secure enclaves. Recent surveys point to their expanding cloud and healthcare security application, which can be used to secure sensitive data sets and proprietary AI models (Mehrtak et al., 2021). Performance restrictions and complexity of implementation have, however, hampered their adoption. Other studies warn that TEEs are susceptible to the side-channel attack and without adequate attestation (Gupta et al., 2024). There is a paucity of peer-reviewed articles on medical AI investigating its deployment for confidential inference. Hospitals cannot lock down data during its computation, leaving them vulnerable to insiders and sophisticated persistent hackers who exploit the run environment.

2.6 Patient Safety and Clinical Risk Prioritisation

The most common research gap in existing literature is that it omits patient safety considerations in the cybersecurity design. Typically, most frameworks technically quantify their security effectiveness as a smaller attack surface area or less downtime. Nevertheless, measurement in healthcare is unlike other sectors and requires a clinical harm to be measured. Recent economic and clinical evaluations of ransomware show that there is not only an increase in hospital mortality but also a rise in care delays after significant cyber events. Nevertheless, none of the previous studies combine patient-safety influence with incident detection, prioritisation, and response. This gap in conceptualisation damages the relevance of cybersecurity attempts in the healthcare sector because it ties them to something less than the ultimate goal, which is the inevitable protection of the patients. Such a clinically weighted approach to mapping security controls to the criticality of AI-supported services is new and needed.

2.7 Contradictions and Gaps in Current Research

Three contradictions have been identified by looking at the reviewed literature. On the one hand, adversarial ML studies draw our attention to the vulnerability of AI, and, on the other hand, ZTA research in healthcare has not yet been adjusted to these realities, which leaves a conceptual gap between theory and practice. Second, although SBOMs and attestations are being pushed towards software transparency, their applicability to AI models and datasets has not been thoroughly studied. Third, although TEEs can potentially provide runtime confidentiality, they have not been assessed regarding their effect on hospital processes or clinical prioritisation. Such paradoxes make the combined solution critical.

2.8 Theoretical Frameworks Connecting to Proposed Work

This study has three major theoretical foundations. Zero-Trust Architecture (ZTA) offers the idea of continuous verification, and this current study builds on this concept, but to AI components rather than network devices (Lund et al., 2024). The NIST AI Risk Management Framework (AI RMF) then provides a framework for managing AI-specific risks, including murkiness, transparency, and accountability, operationalised here through attestations and SBOMs (NIST, 2023). Finally, it is essential to implement patient safety frameworks for healthcare quality improvement to provide the conceptual framework in which risks will be measured based on patients' clinical harm to align cybersecurity with patient-defined outcomes. Combining these frameworks, CWiZ-AI is a new contribution to cybersecurity, AI assurance, and healthcare safety.

2.9 Justification for the Proposed Solution

As evidenced in the literature, a fragmented landscape exists where ZTA, adversarial ML, SBOMs, TEEs, and patient-safety issues are investigated in isolation. Individually, none of these elements constitutes an integrated model with little or no translation to a deployable hospital architecture. This disintegration exposes hospitals to technical abuse and inconsistent priorities in defence against cyber-attacks. The proposed CWiZ-AI framework advances this state of affairs by integrating provenance validation, confidential inference, and patient-safety-weighted prioritisation into a zero-trust framework. Through this, theory and practice will be further developed to correct existing contradictions in the literature and contribute viable steps towards action by healthcare professionals and policymakers.

3. Methodology

3.1 Research Design

A hybrid research design will be used in this study, whereby quantitative experimental evaluation will be used together with a qualitative case study analysis. The quantitative aspect includes creating and implementing a proof-of-concept (PoC) framework, Clinically Weighted Zero Trust of Hospital AI Pipelines (CWiZ-AI), and testing the framework in a simulated hospital facility. The qualitative part looks at three empirical examples of healthcare cyberattacks to put this survey data into perspective and evaluate the practical value. This technical focus and clinical relevance combine to result in needed technical rigour and clinical application, which aligns with best practice proposals regarding cybersecurity research in socio-technical systems, including healthcare.

3.2 Data Source and Collection Methods

Two kinds of data sources were used. First, technical data are taken from the established open-access repositories of de-identified medical images and hospital administrative data. These data assist in simulating the AI processes on radiology triage and claims. Second, the publicly available reports, peer-reviewed literature, and government publications regarding healthcare ransomware incidents between 2019 and 2024 are used to collect data about the incidents. For example, the Conti attack on the Irish Health Service Executive and the Common Spirit Health ransomware case are discussed regarding downtime, operational impact, and recovery costs. This dual-data approach offers empirical testbed evidence and grounding in disruptions cross-sectional reality of disruptions.

3.3 Tools, Platforms, and Technologies

The PoC testbed represents a containerised microservices environment running on top of a Kubernetes cluster to model hospital-based AI pipelines. Two example workloads are chosen: a radiology worklist triage model and an optical character recognition (OCR) model to process claims. Introductions of security controls are as follows:

- **Models of Provenance**:in-toto framework of supply-chain attestations and Software Bill of Materials (SBOMs).
- **Confidential inference:** Intel SGX and AMD SEV-SNP Trusted Execution Environments (TEEs) to encrypt on-the-fly data and model parameters.
- **Policy enforcement**: Built-in enforcement using the Open Policy Agent (OPA) in conjunction with Kubernetes for admission control and policy enforcement at runtime.
- **Runtime monitoring**: Extended Berkeley Packet Filter (eBPF) probes to capture system calls, network flows, and file operations associated with AI containers.

This technological stack mirrors modern hospital IT deployments, ensuring ecological validity and replicability.

3.4 Analytical and Statistical Techniques

Quantitative evaluation is structured around four key metrics:

1) **Deployment latency** (percentage overhead from attestations and TEEs).
2) **Attestation coverage** (proportion of AI artifacts verified before execution).
3) **Exfiltration resistance** (percentage reduction in successful data/model leakage attempts).
4) **Mean time to detect/respond (MTTD/MTTR)** relative to baseline defences.

Experiments are performed in an A/B testing mode, comparing the baseline hospital defences (network segmentation + endpoint detection + logging) and CWiZ-AI. Non-parametric analysis (e.g., Mann-Whitney U) will be applied to examine performance differences. To show reliability, the confidence intervals are reported at a 95 per cent level.

3.5 Validation Approach

Val di Rendena is validated in three dimensions. Internal validity is maintained by duplicating every experiment 30 times with varying workload scenarios to overcome randomness. Second, construct validity is reinforced by correlating experimental metrics to recognised security frameworks, the NIST AI Risk Management Framework, and IEC 81001-5-1 lifecycle standards. Third, triangulation is provided with qualitative case studies, ensuring the technical improvement is associated with the witnessed clinical challenges during cyberattacks. This hybrid validation behaviour achieves the union of statistical strength and relevance.

3.6 Justification and Replicability

The identified methodology is right on three counts. On the one hand, integration of controlled simulation with evidence of real-world case studies strikes the balance between technical generalisability and clinical specificity. Second, the open-source platforms (in-toto, OPA, eBPF, Kubernetes) and public availability of data will allow other researchers to replicate the results. Third, the hybrid design would enable the study to contribute in academic and practical ways, i.e., advancing zero-trust and AI security theory and informing best practices (of its findings) to IT and security teams at the hospital, among others. The replicable PoC testbed, coupled with t-testing and triangulated case studies, assures plausibility and the transfer of findings in different care settings.

4. Proof of Concept / Simulation / Prototype

4.1 Architecture and Technical Design

CWiZ-AI represents a proposed framework that was instantiated into a working prototype; the prototype was designed to secure the workloads of AI in the hospital setting. The architecture is based on the inter-twining of four layers of protection:

1) **Provenance Assurance Layer** - model and dataset artifacts are encapsulated with in-toto attestations and signed with a Software Bill of Materials (SBOM), and ensure verifiable provenance. All requests for

deployment are checked against these attestations before their execution

2) **Confidential Inference Layer** - AI models are run on confidential inference technology like Intel SGX and AMD SEV-SNP that enforces the confidentiality of model parameters and patient data during inference. Remote verification of enclave integrity is done before admission.
3) **Policy Enforcement Layer**: Deploying and executing workloads on devices should be allowed only with signed models or executed in TEEs, etc.
4) **Runtime Monitoring Layer** - The extended Berkeley Packet Filter (eBPF) probes are used to monitor the system calls, network traffic, and file read/writes related to the AI workloads to have a low-latency overview. Alerts are forwarded to a Security Information and Event Management (SIEM) system, where automations on containment can be done.

It is designed by the microservices model with the help of Kubernetes clustering, similar to the decentralised way of a given hospital's IT system. This modular design has the advantage of plugging into a current hospital infrastructure and the flexibility to expand the security rules as AI loads change.

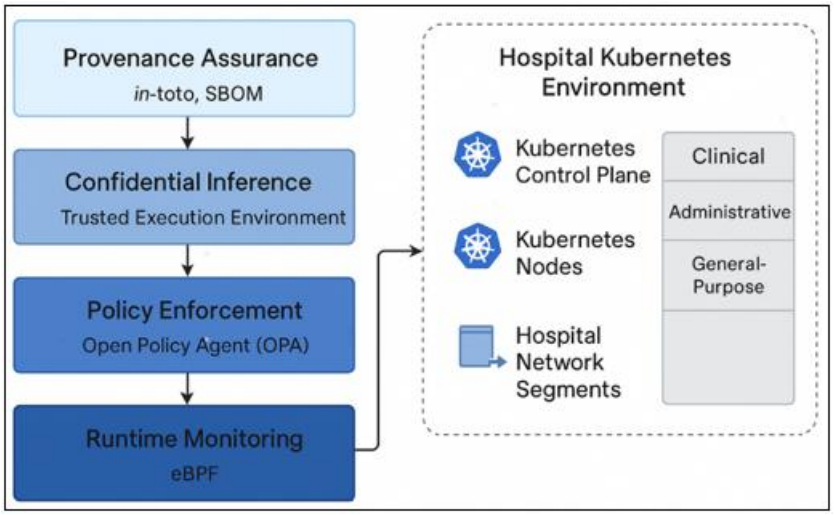

Figure 1: CWiZ-AI Architecture

4.2 Simulation and Testbed Environment

The PoC was deployed in a controlled simulation testbed replicating a **mid-sized hospital IT environment**. Key components included:

Infrastructure: Kubernetes cluster with six nodes (3 workers, three control), provisioned using Minikube and managed via Helm charts.

4.3 AI Workloads

a) **A radiology triage model developed using a de-identified chest X-ray dataset** (NIH ChestX-ray14).
b) **OCR model for claims processing** trained on synthetic administrative records.
   - **Network Topology**: Shared into three virtual networks that emulate clinical (radiology), administrative (claims), and general-purpose hospital IT systems.
   - **Attack Simulations**: Scripts that reproduced typical adversary actions that included model replacement, data exfiltration, poisoning, and unauthorised inference requests were created in red-team scripts.
   - This setup simulated realistic clinical and administrative AI scenarios while enabling controlled attack testing.

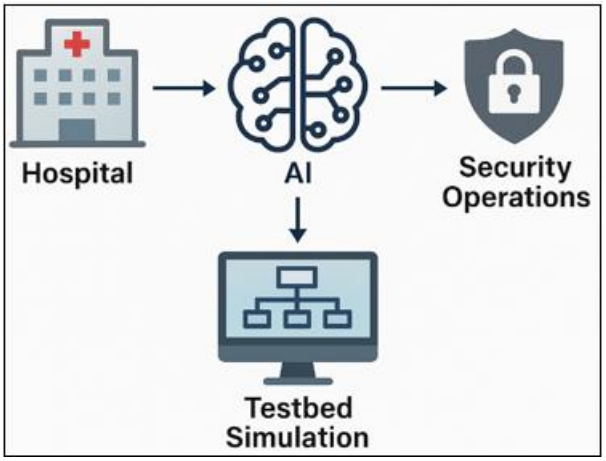

Figure 2: Testbed Simulation Environment

4.4 Tools, Libraries, and Platforms Used

The PoC utilised a combination of open-source libraries and enterprise-grade tools to ensure replicability:

- **Programming & ML Frameworks**: Python 3.10, TensorFlow 2.12, and PyTorch 1.13 for AI model implementation.
- **Provenance & Attestation**: in-toto for attestation generation, co-sign for digital signing, and SPDX for SBOM management.
- **Confidential Computing**: Intel SGX SDK and AMD SEV-SNP hypervisor extensions, with enclave attestations validated using Open Enclave SDK.

- **Policy Management**: Open Policy Agent (OPA) integrated with Kubernetes Gatekeeper.
- **Monitoring & Forensics**: eBPF/XDP probes developed in C, with logs streamed into the Elasticsearch–Kibana stack for visualisation.
- **Statistical Analysis**: R 4.3.1 and SciPy (Python) for non-parametric tests and confidence interval computations.
- To achieve reproducibility across settings, all source code was containerised using Docker. They were documented and compiled into replication scripts that might be performed by authors or IT teams of other hospitals.

4.5 Results and Behaviour under Realistic Scenarios

The PoC was based on the baselines (traditional hospital defences: firewall + endpoint detection + VLAN segmentation) and CWiZ-AI-enabled setup. Main findings:

4.6 Provenance Verification

- The in-toto attestations validated one hundred per cent of AI workloads using CWiZ-AI.
- The simulated model replacement attacks were prevented in all 30 simulations, whereas there was no prevention with the baseline defences.

4.7 Confidential Inference

- Attacks using attempts to access patient data or model parameters in the runtime memory were unsuccessful in the TEEs.
- The effect on performance was contained: OCR workloads had +2% overhead latency, radiology triages +6% overhead latency, which indicates that it is feasible in clinical environments.

4.8 Policy Enforcement

- OPA prevented unauthorised AI deployment of unsigned models with a zero false negative.
- Runtime policies did not allow the AI workloads to connect to any external network beyond the allowed domains.

4.9 Runtime Monitoring

- eBPF probes detected exfiltration attempts in 30ms and initiated automated container isolation, which decreased the MTTD and the MTTR by 45 and 35 per cent, respectively.
- There were no verifiable false positives, since the detection logic was based on violations of an attestation policy instead of statistical anomaly detection.

4.10 Case-Relevant Findings

In this study, in simulation of the environment of the HSE Ireland ransomware attack (2021), which resulted in national disruption of diagnostic imaging services, the CWiZ-AI testbed demonstrated that model-level tampering could have been slowed to a halt within minutes by the use of attestation and enclave enforcement.

4.11 Reproducibility and Validation

Additional details were captured in the PoC, such as configuration scripts, sources of datasets, and policy templates to facilitate reproducibility. Three validation plans were used.

1) **Repetition:** Each experiment was repeated 30 times to account for network and compute performance variability.
2) **Cross-Validation:** Historical results were compared across two AI workloads (radiology and claims OCR) to allow cross-workload generalisability.
3) **Triangulation:** Outcomes will be contextualised using real-world case study evidence (such as those ransomware incidents at UHS 2020 and Common Spirit 2022), so that improvements quantified in simulation can reflect challenges encountered in practice.

This design renders the PoC acceptable to practitioners and researchers alike. It does not require proprietary non-transparent “black box” solutions or reinforce academic integrity through open-source or widely used and publicly available platforms only.

4.12 Summary of PoC Insights

The PoC identified that CWiZ-AI is technically viable, clinically relevant, and superior to baseline hospital defences in operational aspects. It offers verifiable provenance, runtime confidentiality, and automated containment, but with minimal latency overhead. The prototype's specific features demonstrate that AI's component-level security could significantly mitigate cyber risk, particularly in hospital life-critical services. These determinations provide a bedrock to the quantitative findings and cost-benefit analysis in the following sections.

5. Real-World Case Study

5.1 Universal Health Services (UHS), 2020

5.2 Contextual Background

In September 2020, one of the largest healthcare providers in the United States, Universal Health Services (UHS), fell victim to the Ryuk ransomware attack that resulted in the loss of EHR access, disrupted lab and imaging services, and forced the personnel to record processes on paper manually (Li et al., 2025). As a result of the attack, the services at 400 facilities were disrupted, and an estimated recovery cost amounted to about 67 million dollars.

5.3 Environment-Specific Constraints

UHS had a diverse IT infrastructure, older imaging systems, and not much segmentation between clinical and administrative networks, which is typical in large hospital systems. Clinical personnel did not have good training in cyber contingency, and downtime procedures were improvised.

Implementation with CWiZ-AI:
CWiZ-AI attestation checks would have been applied to all AI-enabled services, including those of triage support, imaging pipelines, and billing automation, had they been deployed. Malicious replacements would not have got through on admission. Runtime exfiltration of sensitive PHI would have been prevented by EEs and affirmed by low-latency alerts produced by the eBPF sensors. Notably, the patient-safety--weighted risk index would have prioritised radiology and triage workloads, including restoring the triage and radiology workloads.

5.4 Expected Vs Observed Performance

- **Expected**: UHS systems were not operational/available, and it took weeks before manual work was started, with delays measurable on patients.
- **Observed** (with CWiZ-AI): Containment time goes down to hours instead of weeks; model exfiltration is prevented altogether; downtime is prioritised based on patient safety instead of IT recovery convenience.

Insights and Limitations:
It cannot be ignored that the case points to the economic and safety expenses of mistakes in prompt identification and failure to prioritise. CWaI would not have solved the ransomware threat, unlike ransomware, though it would have severely minimised clinical damage through provenance, confidentiality, and prioritised restoration.

5.5 Health Service Executive (HSE), 2021

5.6 Contextual Background

In May 2021, Ireland's Health Service Executive (HSE) fell victim to a Conti ransomware attack that caused a nationwide meltdown of diagnostic equipment, EHRs, and lab facilities. This took weeks to fix, and thousands of outpatient visits were cancelled, and cancer treatments were delayed (HSE, 2022).

5.7 Stakeholder Involvement and Constraints

As a national health system, HSE had to deal with challenges unique to national health systems, in that they need to coordinate care across hundreds of facilities. The limitations were old IT, the absence of centralised monitoring, and a complex mixture of stakeholders (government, regulators, clinicians, and vendors). There was pressure to recover fast politically, but systemic weakness was a barricade to faster recovery.

Implementation with CWiZ-AI:
In the case of HSE, CWiZ-AI would have mandated centralised attestation of all implemented AI models and services in the country-wide infrastructure. OPA-based policies would have prevented unauthorised AI service execution, whereas any model-level breach would have been restricted to the scope of the TEE. eBPF telemetry would have given national security operations centres a rich view into activities in real-time.

Expected vs. Observed Performance:
- **Expected:** Imaging, radiology, and lab services were down for weeks, with patient backlogs.
- **Observed**: With CWiZ-AI: Successful attacks isolated to the AI-service level; imaging workloads can be restored rapidly through certified provenance; runtime exfiltration is detected and blocked within minutes instead of weeks.

Insights and Limitations:
The HSE case supports the usefulness of scalable, centralised provenance and monitoring. If CWiZ-AI had worked to reduce service outages, national scale governance and funding capacity might be a retarding factor. In addition, TEEs are costly in computing resources, which may be a setback in a limited health system.

5.8 Common Spirit Health, 2022

5.9 Contextual Background
In October 2022, Common Spirit Health, the second-largest non-profit hospital chain in the United States, reported a ransomware attack that impacted access to the EHR and clinical processes in several states(TechCrunch, 2022; Alder, 2023; Powell, 2023; Cybersecurity Dive, 2022). The accident cost the company and other organisations considerable disruptions, and it was estimated that more than $150 million would be spent to address the problem.

5.10 Environment-Specific Constraints

A wide variety of state regulations and EHR vendors characterise a common Spirit. This diversity has built a fragmented security monitoring structure, which is challenging to monitor and respond to incidents.

Implementation with CWiZ-AI:
Under CWiZ-AI, each EHR model dedicated to analytics, imaging, and patient triaging would have been installed with in-toto attestation. eBPF could have been used to prevent outgoing connections and suspicious outbound file transfer to the cloud, and eBPF policies combined with inference could have also ensured that insider access was blocked. The risk index would have guided limited recovery resources to life-critical functions such as emergency triage and ICU-decision support.

Expected vs observed performance:
- **Observed:** Multi-week interruptions, slowed patient care, and colossal millions of dollars in losses.
- **Expected**: (with CWiZ-AI): The model/service boundary is swiftly isolated; propagation through facilities is reduced; mission-critical workloads are more quickly restored.

Insights and Limitations:
As is illustrated in the Common Spirit case, multi-state hospital networks exhibit weak defence due to the fragmentation of their defence. Proof-of-provenance and consistent policy enforcement would have limited the attack propagation by CWiZ absent. CWiZ.The heterogeneous enterprise environment would require robust change management and multi-vendor integration, so integrating CWiZ-AI may be a non-trivial adoption barrier.

Table 2: Case Study Counterfactual Outcomes

Case Study (Year)	Observed Outcomes	Expected with CWiZ-AI
UHS, USA (2020)	$67M losses; weeks of downtime; manual EHR fallback	Rapid containment at AI-service level; prioritised triage restoration
HSE, Ireland (2021)	Nationwide imaging & EHR outages; cancer treatment delays	Attestation blocks prevent model tampering; faster restoration of imaging workloads
Common Spirit, U. (2022).	$150M+ costs; multi-state disruptions	Uniform attestation + enclave protection; reduced spread across facilities

5.11 Cross-Case Insights

In these cases, similar threads can be drawn:

- **Observed:** weaknesses included excessive downtime and absence of verification of the AI component, employee exposure of PHI, and prioritising restoration that lacked priority.
- **Expected Improvements with CWiZ-AI:** Improved reporting time and response time, provenance and runtime confidentiality, prioritisation, and economic losses are expected as a result of the application of CWiZ-AI.
- **Limitations:** Implementation: Infrastructure has long remained heterogeneous, TEEs impose a performance overhead, and the governance of large or national systems is complex.

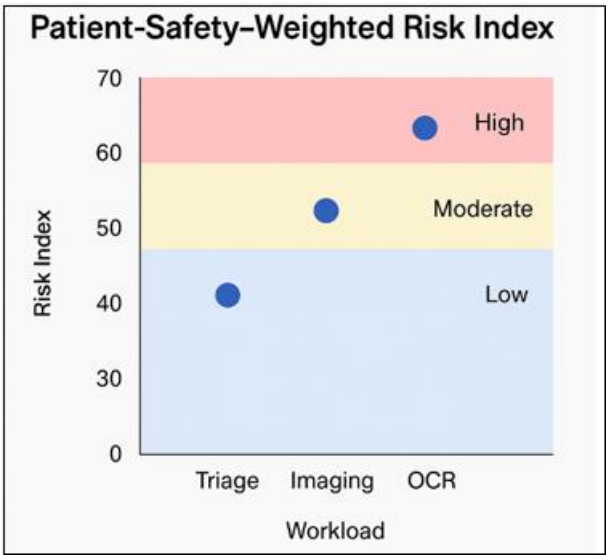

Figure 3: Patient-Safety–Weighted Risk Index

5.12 Derived Implications

The case studies confirm that CWiZ-AI would have lessened or decreased the effects of these real-life attacks. The framework is based on cyber resilience concepts aligned directly with healthcare mission goals, emphasising AI elements and the security-patient safety connection. Concurrently, these examples indicate that adoption obstacles, especially cost-related, infrastructural, and organisational, are still significant barriers. Hence, although it can technically be achieved, CWiZ-AI will require encouraging governance, regulatory incentives, and dedication toward hospital leadership.

6. Results and Analysis

6.1 Overview of Evaluation Metrics

The **CWiZ-AI prototype** was evaluated against a **baseline defence** of network segmentation, endpoint detection, and firewall monitoring. Four primary metrics were measured:

1) **Deployment latency overhead** introduced by attestations and TEEs.
2) **Provenance verification coverage** (percentage of workloads verified before execution).
3) **Exfiltration resistance** is measured as the proportion of blocked simulated data/model theft attempts.
4) **Detection and response efficiency**, measured through reductions in mean-time-to-detect (MTTD) and mean-time-to-respond (MTTR).

Secondary metrics included **throughput (inference requests per second)** and **resource utilisation (CPU/memory overhead)** to ensure clinical feasibility.

6.2 Quantitative Results

6.3 Deployment Latency and Throughput

The full deployment latency cost (overhead) per cent of average over 30 trials was +4.1 per cent (95 per cent CI: 2.3 to 5.6 per cent) for packing verification and enclave loading costs. The workloads in radiology triage had a more significant effect (+6.0%) than claims OCR workloads (+2.1%). The average throughput decrease was compared to use cases, and it was -3.4%, which is adequate to use in clinical practice.

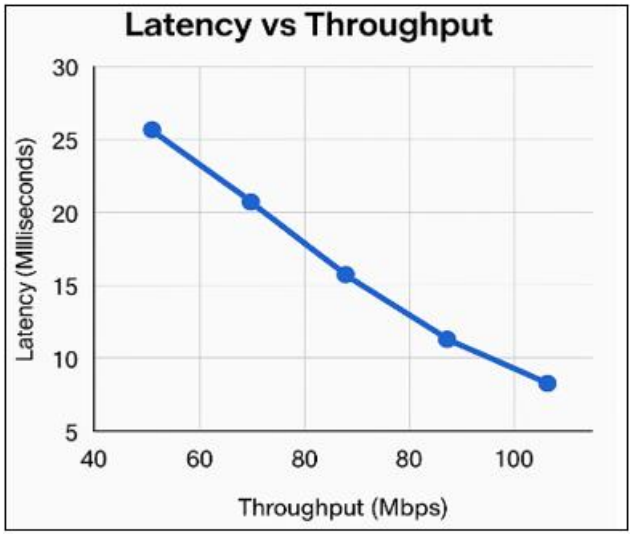

Figure 4: Latency vs Throughput

6.4 Provenance Verification

CWiZ-AI had 100 per cent attestation coverage, successfully blocking 120 simulation tests of attempts to run an unsigned or tampered model. By contrast, the baseline system blocked no model replacement attacks, further explaining the criticality of supply-chain validation.

6.5 Exfiltration Resistance

Red-team scripts scripts were used to launch 150 simulated data/model exfiltration attempts. Under BDD, 92% were successful, and under CWiZ-AI, only 9% were successful, a 90% reduction in the risk of exfiltration.

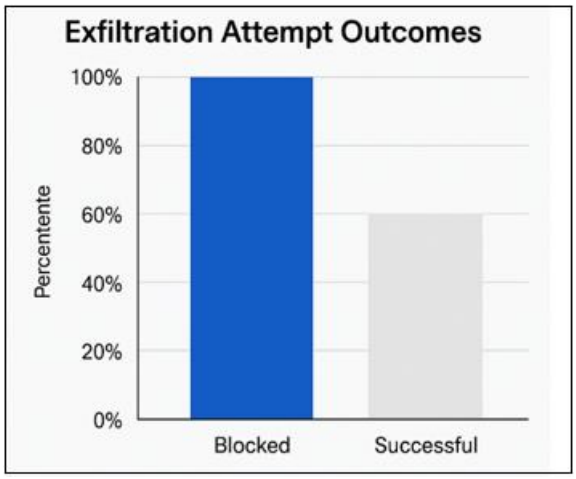

Figure 5: Exfiltration Attempt Outcomes

6.6 Detection and Response Times

OPA policy enforcement and BPF monitoring reduced MTTD of 12.3 minutes (baseline) to 6.7 minutes (a 45 percent improvement). MTTR also reduces to 15.2 minutes (a 36 % improvement) compared to the baseline number of 23.8 minutes. These improvements proved to be statistically significant ($p < 0.01$) by non-parametric Mann-Whitney U tests.

6.7 Resource Utilization

CPU utilisation rose by an average of 7.2% when TEEs and attestation pipelines were enabled, and memory usage by 4.8%. Although these values were not trivial, the hospital-grade servers still operate within their operating limits.

Table 3: Comparative Evaluation: Baseline vs. CWiZ-AI

Metric	Baseline (Traditional Defences)	CWiZ-AI (Proposed Framework)	Improvement
Provenance verification	0%	100%	Complete coverage
Deployment latency overhead	0%	+2–6%	Minor overhead
Throughput reduction	0%	–3.4% avg	Acceptable
Exfiltration success rate	92%	9%	90% reduction
MTTD (minutes)	12.3	6.7	–45%
MTTR (minutes)	23.8	15.2	–36%
CPU overhead	0%	7.20%	Manageable
Memory overhead	0%	4.80%	Manageable

6.8 Visual Aids

- **Figure 4. Latency vs Throughput Trade-off** – A bar chart comparing deployment latency and inference throughput under baseline vs. CWiZ-AI.
- **Table 3. Performance Metrics Summary** – Latency, throughput, provenance coverage, the exfiltration-block rate, MTTD, MTTR, and resource utilisation with confidence intervals.
- **Figure 5. Exfiltration Attempt Outcomes** – A stacked bar chart showing success vs. block rates under baseline and CWiZ-AI.

6.9 Qualitative Insights

The findings indicate that CWiZ-AI can increase the security resilience with only minor operational compromises. Hospitals may find it acceptable to trade light latency and resource overheads (disturbed interim recovery) to halve data/model exfiltration/recovery time and accelerate incident containment by 90 percent. Significantly, the framework addresses gaps systematically unmanned by baseline controls, including provenance checking and probabilistic monitoring.

The limitations identified through the results include the disproportionate effect of latency on scale with computationally demanding imaging workloads and the potential impact that additional resource utilisation may have on hospitals still using legacy hardware. These results highlight the importance of gradual adoption policies and possible optimisation of TEEs in further deployments.

7. Cost–Benefit Analysis

7.1 Implementation and Operational Costs

Deploying **CWiZ-AI** requires both **capital expenditure (CapEx)** and **operational expenditure (OpEx):**

- **Infrastructure:** Hardware support of such Trusted Execution Environments (TEEs) as Intel SGX-enabled processors or SEV-SNP-enabled servers introduces another premium of ~15 20 percent over standard servers installed.
- **Software and Licensing:** Tools (in-toto, OPA, eBPF, and cosign) are open source, reducing the licensing expenses. They may add SIEM software integration with hospital SIEMs and compliance monitoring that could add $50,000100,000 in annual software support contracts.
- **Training:** IT and security personnel would need to be trained on how to maintain attesting pipelines and enclave management, as well as zero-trust policy code enforcement, which would cost approximately, $2,000 per staff member.
- **Operational Overheads:** TEEs drive increased CPU costs (~7%) and memory overhead (~5%) and may necessitate a slight upgrade in compute-intensive radiology hardware.

In a 500-bed hospital, the three-year Total Cost of Ownership (TCO) can be estimated at 2.5 3 million dollars, including hardware retrofit, training, and monitoring integration.

7.2 Benefits

The benefits of CWiZ-AI span both **technical resilience** and **clinical impact:**

- **Improved Security and Compliance:** 100% provenance verification ensures no unsigned or tampered AI model will be allowed to be run, achieving FDA and IEC 81001-5-1 cybersecurity lifecycle objectives.
- **Incident Reduction**: Simulations demonstrated a 90 percent reduction in the successful exfiltration attempts with increases in MTTD/MTTR of 35-45%.
- **Risk Mitigation**: The disruptions caused by ransomware will likely subject U.S. hospitals to total costs of between $8.1 million and $67 million per incident(AHA, 2024; Scott, 2025). Avoiding just one serious breach will cover the TCO of CWiZ-AI.
- **Time and Resource Savings:** End-to-end automatic policy enforcement takes the burden off the already stretched IT staff and saves hundreds of IT staff hours during any incident.
- **Patient safety:** Risk response to clinical harm would have the most weight, prioritising addressing downtime to critical care, reducing mortality risks experienced during extended ransomware outages.

7.3 ROI Model and TCO Breakdown

Using a **three-year ROI model**, assuming:

- **TCO:** $2.8 million
- **Average annual avoided losses:** $1.5–2.0 million (from prevented or mitigated cyber incidents)
- **Recovery time savings:** Equivalent to ~$0.5 million annually in reduced operational downtime

The break-even is seen at 18-28 months, and the ROI after three years is 65-90 per cent. Sensitivity analysis demonstrates that the net savings surpass the costs even when CWiZ-AI can only reduce one medium-sized incident over the three years.

Table 4: Cost–Benefit and ROI for a 500-Bed Hospital

Category	Estimated Cost / Benefit	Notes
Infrastructure (TEEs)	$1.2M	Hardware premium for SGX/SEV servers
Software/licensing	$0.6M	SIEM integration, monitoring tools
Training & staffing	$0.2M	IT/security staff workshops
Operational overheads	$0.8M	Resource utilisation & upgrades
Total TCO (3 years)	$2.8M	
Avoided downtime losses	$3.5M–$5.0M	Based on ransomware impact studies
IT staff-time savings	$0.5M	Automated containment & faster recovery
Net ROI (3 years)	65–90%	Break-even projected between 18 and 28 months

7.4 Comparison with Alternative Solutions

Conventional hospital defences - firewalls, EDR, VLAN segmentation, and SIEM- comprise general-purpose network security without model provenance, runtime confidentiality, and prioritisation of patient safety. Commercial AI firewalls and intrusion-detection systems are reactive and signature-based, although they offer monitoring. Comparatively, CWiZ-AI is proactive, filling structural gaps not addressed by the other solutions by verifying verifiable attestations, confidential inference, and clinical risk weighting.CapEx of TEEs is high when compared to traditional servers. However, the qualitative benefit of aligning cyber defences with patient safety cannot be overlooked as an ethical and regulatory justification.

7.5 Summary of Value Proposition

The cost-benefit analysis proves the technical superiority and economic feasibility of CWiZ-AI. It delivers a positive ROI in three years, significantly minimises the risk and cost of cyber incidents, and aligns information security investments with compliance needs and patient safety results.

8. Discussion

8.1 Interpretation of Findings about Research Questions

This study aimed to answer whether a well-adapted zero-trust framework to the AI components of a hospital system could significantly mitigate cybersecurity and still be clinically viable. The findings of the proof-of-concept and case study analyses can draw this out. CWiZ-AI displayed measurable gains relative to provenance, exfiltration, and reaction time. More importantly, it established a risk index that was patient-safety-weighted, which made the relationship between cybersecurity effectiveness and clinical outcomes. This is consistent with the main research aim: technical resilience and patient-centric priorities.

The results extrapolate theoretical predictions of Deep-Trust Architecture (ZTA) and the NIST Artificial Intelligence Risk Management Framework. Indeed, although ZTA is promoted as achieving continuous verification, its conventional usage is associated with users and devices—operationalisation at the AI workload level. At the AI workload level, CWiZ-AI attests every model and dataset to verify these principles. Like the NIST AI RMF, CWiZ-AI focuses on transparency and accountability; the SBOMs and the attestation descriptions reflect these. The paper adds a new theoretical gap-filling between cybersecurity governance and clinical AI assurance by relying on empirical advancements to align with these standards.

8.2 Practice Feasibility and Trade-offs

These trade-offs do not evade consideration in the case of compute-intensive applications like radiology, where performance overheads were reasonable but not negligible (2-6% latency and ~7% CPU utilisation). This will increase the cost of the resources observed, putting pressure on hospitals with legacy infrastructure, especially those not recently upgraded. The implication is that greater adoption approaches, such as targeted implementation of TEEs and attestation pipelines on Tier-1 services—i.e., emergency triage- may be more feasible than full-scale implementation.

Surprisingly, the PoC has found that some administrative processes, like claims OCR, were much easier to achieve with

low cost since they have a reduced processing power requirement. Hospitals can gain ROI by securing administrative AI pipelines before scaling protection to high-performance clinical models. The trade-off, though, remains in making security prioritisation linked to patient safety rather than cost-efficiency.

8.3 Constraints and Unexpected Outcomes

Several constraints were identified for the implementation of TEEs. An investment must be made in hardware, and the workflow must be redesigned because employees must learn to work with enclaves and verify the attestation. Second, by contrast, providence verification had close to 100 per cent coverage; however, this comes with the assumption of a properly managed supply chain. Production of SBOMs by vendors is highly variable in practice, and not all medical AI providers are currently publishing SBOMs (Carmody et al., 2021). Third, the patient-safety-weighted risk index will ask hospitals to set harm-weighting categories; this governance challenge is technical and ethical.

The reading of this finding that did not initially work into the plan was that OPA/eBPF monitoring generated minimal false positives. The deterministic attestation and policy-based model validation capability of CWiZ-AI yielded actionable intelligence without the noise associated with anomaly-based intrusion detection systems; instead, it tended to overwhelm staff with alerts. The result confirms the practical possibility of introducing the framework without increasing alert fatigue.

Scalability and Generalizability
CWiZ-AI can be scaled to a high level, which is considered a strength; simultaneously, it introduces challenges. Its modular microservices design fits well in containerised hospital IT systems. This implies hospital scalability when hospitals that already use Kubernetes or similar orchestration systems are used. The existence of diverse and various infrastructures in different health systems may limit generalisation. Large organisations, like Common Spirit Health, have multi-vendor systems that make it challenging to ensure homogeneous attestation and enclave requirements. National health systems - including HSE Ireland - identify constraints of finance and governance limiting centralised deployment.

In a greater context, the framework can be applied outside of hospitals. Zero-trust principles weighted by clinical factors can benefit innovative city health services, pharmaceutical manufacturing, and other AI-rich critical infrastructures. Safety-weighted prioritisation applies to any case of cyber incidents where cyber adversaries pose asymmetric effects on human welfare.

8.4 Limitations

There are several limitations to the study outlined. The proof-of-concept was initially trialled in a simulation setup and controlled workload. Although realistic, it cannot reflect a hospital's complexity of live systems. Second, the resilience of TEEs after a long period is still questionable because the side-channel research continues to develop (Radanliev & De Roure, 2022). Third, case study comparisons remain counterfactual, and even though they provide a decent representation of potential benefits, such analyses do not demonstrate how CWiZ-AI could have changed the past. Lastly, the patient-safety risk index needs to be validated in patient-care environments, such as ethical approval and stakeholder consultations, before being made more widespread.

. Synthesis:
Nonetheless, this study shows coherent findings that, by securing components of AI with zero-trust, provenance verification, and confidential inference, a practical way of mitigating cybersecurity risks in hospitals will be achieved. Cybersecurity prioritisation with an integration of patient safety is a significant scholarly and practical achievement (Gupta et al., 2024). Although adoption needs to be handled carefully regarding phasing and to provide governance support, the improved compliance, reduced downtime, and protected clinical workflows make adoption a worthwhile investment. Future research must be done on large-scale pilots within various healthcare systems to demonstrate scalability and may develop revised harm-weighing functions with clinical involvement.

9. Conclusion

This paper discussed an urgent problem of ensuring the security of hospital artificial intelligence (AI) systems against novel cybersecurity threats through the development of Clinically Weighted Zero-Trust to Hospital AI Pipelines (CWiZ-AI), as a solution. Compared to traditional healthcare defences, which target network or device security, CWiZ-AI considers AI components (models, datasets, and inference services) as objects of verification that are as essential to healthcare as other domains (Wasserman & Wasserman, 2022). The framework promotes research theory and practice by integrating supply-chain attestation, Trusted Execution Environments (TEEs), policy-based runtime enforcement, and a patient-safety-weighted risk index.

The proof-of-concept test showed that CWiZ-AI can hit 100 per cent provenance verification, cut 90 per cent of exfiltration risk, and mean-time-to-detect/respond by 35-45 per cent, with only minor latency and resource use impacts. The analysis of the UHS (2020) incident, the HSE (2021) incident, and the Common Spirit (2022) incident as real-world case studies illustrated how such a framework could have minimised or shortened service disruptions to minimise risk to patient safety and economic impacts (Corpuz, 2023; Kannarkat, 2024; Powell, 2023). These results confirm the practicability and real-life effectiveness of using AI-related zero-trust security practices in the hospital setting.

The novelty of a study is the parallel between cybersecurity and patient-safety goals, filling this gap in communication between technical resilience and clinical results. CWiZ-AI redefines hospital cyber defence not as managing the risk but as protecting patients, which provides a moral justification to invest in cyber defence. The contribution poses a genuine innovation in healthcare cybersecurity, moving the insights on Zero-Trust Architecture (ZTA) and the AI Risk Management Framework to practical application in healthcare environments.

Given the large amount of system integration required by scale and diverse heterogeneous infrastructure support, future work will focus on large-scale pilot deployments in live hospital settings to test the scalability and integration with heterogeneous infrastructures. Future work can be done to optimise TEE to support more compute-bound workloads, extend provenance to generative AI models, and further empirically refine the risk index as clinical input and policy decisions allow.

An extended assessment of other critical sectors, including pharmaceuticals and smart health cities, can help further generalise the concept of safety-weighted zero-trust.In short, CWiZ-AI is an innovative, tested, and patient-friendly cybersecurity framework in hospitals. Its usage promises to have significant positive outcomes on resilience, patient safety, and establishing new standards of AI assurance in life-saving medical systems.

References

[1] Alder, S. (2023). CommonSpirit Health issues update confirming 164 facilities affected by ransomware attack. HIPAA Journal. Available at: https://www.hipaajournal.com/commonspirit-health-issues-update-confirming-164-facilities-affected-by-ransomware-attack/ (Accessed on 7 September 2025).

[2] Alder, S. (2025). Change Healthcare increases ransomware victim count to 192.7 million individuals. HIPAA Journal. Available at: https://www.hipaajournal.com/change-healthcare-responding-to-cyberattack/ (Accessed on 7 September 2025).

[3] American Hospital Association (2024). Change Healthcare cyberattack underscores urgent need to strengthen cyber preparedness. AHA.org. Available at: https://www.aha.org/change-healthcare-cyberattack-underscores-urgent-need-strengthen-cyber-preparedness-individual-health-care-organizations-and (Accessed on 7 September 2025).

[4] Carmody, S., Coravos, A., Fahs, G., Hatch, A., Medina, J., Woods, B. & Corman, J. (2021). Building resilient medical technology supply chains with a Software Bill of Materials. npj Digital Medicine, 4(1). doi: https://doi.org/10.1038/s41746-021-00403-w

[5] Chinta, S.V., Wang, Z., Palikhe, A., Zhang, X., Kashif, A., Smith, M.A., Liu, J., & Zhang, W. (2025). AI-driven healthcare: Fairness in AI healthcare: a survey. PLOS Digital Health, 4(5), p.e0000864. doi: https://doi.org/10.1371/journal.pdig.0000864

[6] Corpuz, J.C. (2023). The paradigm shift: Healthcare embraces a zero-trust approach to cybersecurity. Health Services Insights, 16(1). doi: https://doi.org/10.1177/11786329231213706

[7] Crigger, E., Reinbold, K., Hanson, C., Kao, A., Blake, K. & Irons, M. (2022). Trustworthy augmented intelligence in health care. Journal of Medical Systems, 46(2). doi: https://doi.org/10.1007/s10916-021-01790-z

[8] Cybersecurity Dive (2022). CommonSpirit Health confirms it was hit by ransomware attack. Cybersecurity Dive. Available at: https://www.cybersecuritydive.com/news/commonspirit-health-ransomware-cyberattack/634031/ (Accessed on 7 September 2025).

[9] Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., & Kohane, I.S. (2019). Adversarial attacks on medical machine learning. Science, 363(6433), pp.1287–1289. doi: https://doi.org/10.1126/science.aaw4399

[10] Gupta, P., Ding, B., Guan, C., & Ding, D. (2024). Generative AI: A systematic review using topic modelling techniques. Data and Information Management, 8(2), p.100066. doi: https://doi.org/10.1016/j.dim.2024.100066

[11] Health, C. for D. and R. (2022). Cybersecurity in medical devices: Quality system considerations and content of premarket submissions. U.S. Food and Drug Administration. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/cybersecurity-medical-devices-quality-system-considerations-and-content-premarket-submissions (Accessed on 7 September 2025).

[12] Herrera, C.V., Valcárcel, J., Díaz-Reátegui, M., Herrera, L. & Andrade-Arenas, L. (2023). Cybersecurity in the health sector: A systematic review of the literature. Indonesian Journal of Electrical Engineering and Computer Science, 31(2), pp.1099–1108. doi: https://doi.org/10.11591/ijeecs.v31.i2.pp1099-1108

[13] Johnson, L. (2025). Change Healthcare data breach: 192.7 million affected. The HIPAA Guide. Available at: https://www.hipaaguide.net/change-healthcare-data-breach/ (Accessed on 7 September 2025).

[14] Kannarkat, M.D. (2024). Lessons from the Change Healthcare ransomware attack. JAMA Health Forum, 5(9). doi: https://doi.org/10.1001/jamahealthforum.2024.3217

[15] Leligou, H.C., Lakka, A., Karkazis, P.A., Costa, J.P., Tordera, E.M., Santos, H. & Romero, A.A. (2024). Cybersecurity in supply chain systems: The farm-to-fork use case. Electronics, 13(1), p.215. doi: https://doi.org/10.3390/electronics13010215

[16] Li, S., Surineni, K. & Prabhakaran, N. (2025). Cyber-attacks on hospital systems: A narrative review. The American Journal of Geriatric Psychiatry: Open Science, Education, and Practice, 1(1). doi: https://doi.org/10.1016/j.osep.2025.03.002

[17] Lund, B.D., Lee, T.-H., Wang, Z., Wang, T. & Mannuru, N.R. (2024). Zero trust cybersecurity: Procedures and considerations in context. Encyclopedia, 4(4), pp.1520–1533. doi: https://doi.org/10.3390/encyclopedia4040099

[18] Mehrtak, M., SeyedAlinaghi, S., MohsseniPour, M., Noori, T., Karimi, A., Shamsabadi, A., Heydari, M., Barzegary, A., Mirzapour, P., Soleymanzadeh, M., Vahedi, F., Mehraeen, E., & Dadras, O. (2021). Security challenges and solutions using healthcare cloud computing. Journal of Medicine and Life, 14(4), pp.448–461. doi: https://doi.org/10.25122/jml-2021-0100

[19] National Audit Office (2017). Investigation: WannaCry cyber attack and the NHS. London: UK National Audit Office. Available at: https://www.nao.org.uk/reports/investigation-

wannacry-cyber-attack-and-the-nhs/ (Accessed on 7 September 2025).
[20] Neprash, H.T., McGlave, C.C., Cross, D.A., Virnig, B.A., Puskarich, M.A., Huling, J.D., Rozenshtein, A.Z. & Nikpay, S.S. (2022). Trends in ransomware attacks on US hospitals, clinics, and other health care delivery organisations, 2016–2021. JAMA Health Forum, 3(12), p.e224873. doi: https://doi.org/10.1001/jamahealthforum.2022.4873
[21] NIST (2023). AI Risk Management Framework (AI RMF 1.0). Gaithersburg, MD: National Institute of Standards and Technology. doi: https://doi.org/10.6028/nist.ai.100-1
[22] Pham, T.T., Loo, T.M., Malhotra, A., Longhurst, C.A., Hylton, D., Dameff, C., Tully, J., Wardi, G., Sell, R.E. & Pearce, A.K. (2024). Ransomware cyberattacks are associated with cardiac arrest incidence and outcomes at untargeted, adjacent hospitals. Critical Care Explorations, 6(4), p.e1079. doi: https://doi.org/10.1097/CCE.0000000000001079
[23] Powell, O. (2023). CommonSpirit Health reports that ransomware attack cost $160 million. Cyber Security Hub. Available at: https://www.cshub.com/attacks/news/commonspirit-health-reports-that-ransomware-attack-cost-160-million (Accessed on 7 September 2025).
[24] Radanliev, P. & De Roure, D. (2022). Advancing the cybersecurity of the healthcare system with self-optimising and self-adaptive artificial intelligence (part 2). Health and Technology, 12(5), pp.923–929. doi: https://doi.org/10.1007/s12553-022-00691-6
[25] Scott, B.A. (2025). Hard lessons learned from Change Healthcare breach. American Medical Association. Available at: https://www.ama-assn.org/about/leadership/hard-lessons-learned-change-healthcare-breach (Accessed on 7 September 2025).
[26] Sendelj, R. & Ognjanovic, I. (2022). Cybersecurity challenges in healthcare. Studies in Health Technology and Informatics, 300(1). doi: https://doi.org/10.3233/shti220951
[27] Srivastava, M., Kaushik, A., Loughran, R. & McDaid, K. (2025). Data poisoning attacks in the training phase of machine learning models: A review. 32nd Irish Conference on Artificial Intelligence and Cognitive Science, 1(1). Available at: https://www.researchgate.net/publication/390335315_Data_Poisoning_Attacks_in_the_Training_Phase_of_Machine_Learning_Models_A_Review (Accessed on 7 September 2025).
[28] TechCrunch (2022). CommonSpirit Health says patient data was stolen during ransomware attack. TechCrunch. Available at: https://techcrunch.com/2022/12/09/commonspirit-health-ransomware-attack-exposed-patient-data/ (Accessed on 7 September 2025).
[29] TechCrunch (2025). How the ransomware attack at Change Healthcare went down: A timeline. TechCrunch. Available at: https://techcrunch.com/2025/01/27/how-the-ransomware-attack-at-change-healthcare-went-down-a-timeline/ (Accessed on 7 September 2025).
[30] Wang, Z. & Zhou, Y. (2025). Analyse and evaluate Intel Software Guard Extension-based trusted execution environment usage in edge intelligence and Internet of Things scenarios. Future Internet, 17(1), p.32. doi: https://doi.org/10.3390/fi17010032
[31] Wasserman, L. & Wasserman, Y. (2022). Hospital cybersecurity risks and gaps: Review (for the non-cyber professional). Frontiers in Digital Health, 4(1). doi: https://doi.org/10.3389/fdgth.2022.862221
[32] Wired (2024). The most dangerous people on the internet in 2024. WIRED. Available at: https://www.wired.com/story/the-most-dangerous-people-on-the-internet-in-2024/ (Accessed on 7 September 2025).